The universal ancestor was a thermophile or a hyperthermophile

Massimo Di Giulio*

International Institute of Genetics and Biophysics, CNR, Via G. Marconi 10, 80125 Naples, Italy

Received 28 September 2001; accepted 5 November 2001

Received by G. Bernardi

Abstract

By exploiting the correlation between the optimal growth temperature of organisms and a thermophily index based on the propensity of amino acids to enter thermophile/hyperthermophile proteins, an analysis is conducted in order to establish whether the last universal common ancestor (LUCA) was a mesophile or a (hyper)thermophile. This objective is reached by using maximum parsimony and maximum likelihood to reconstruct the ancestral sequences of the LUCA for two pairs of sets of paralogous protein sequences by means of the phylogenetic tree topology derived from the small subunit ribosomal RNA, even if this is rooted in all three possible ways. The thermophily index of all the reconstructed ancestral sequences of the LUCA belongs to the set of the thermophile/hyperthermophile sequences, thus supporting the hypotheses that see the LUCA as a thermophile or a hyperthermophile.

Keywords: Paralogous protein; Thermophily index; Ancestral sequence; Last universal common ancestor; Origin of life

1. Introduction

Galtier et al. (1999) introduce a simple idea that makes it possible to differentiate between the mesophilic and thermophilic nature of the last universal common ancestor (LUCA). By exploiting (Galtier et al., 1999) the correlation between the optimal growth temperature of prokaryotes and the G + C content of ribosomal RNAs (Galtier and Lobry, 1997), and estimating the G + C content of the LUCA’s ancestral sequence, Galtier et al. (1999) are able to establish whether the LUCA was a mesophile or a thermophile by noting whether this content lies between the mesophilic or thermophilic sequences. Clearly, if we had an equivalent relationship between the optimal growth temperature of organisms and a variable derived from the amino acid composition of proteins, then the idea of Galtier et al. (1999) could be extended to any type of protein. Such a variable does in fact exist (Di Giulio, 2000a) and is based on the differing propensity of amino acids to enter mesophile or thermophile/hyperthermophile proteins (Di Giulio, 2000a). Therefore, in the present paper, I use the strong correlation between the optimal growth temperature of organisms and a thermophily index (Di Giulio, 2000a) and employ methods for the reconstruction of the LUCA’s ancestral sequence based on maximum parsimony and maximum likelihood to establish whether the LUCA was a mesophile or a (hyper)thermophile.

2. Materials and methods

The signal recognition particle (SRP) sequences of both the 54 kDa and the subunit α (FtsY) were obtained from the web site http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html. The set of sequences was aligned using the CLUSTALX program (Thompson et al., 1997). The alignment of the sequences of these two paralogous proteins is equal, with the exception of a handful of sites, to the one reported in Gribaldo and Cammarano (1998). All the sites containing at least one gap were eliminated from the alignment, which turned out to be 220 residues long. This elimination was necessary because the algorithm which reconstructs the ancestral sequences by means of maximum likelihood (Zhang and Nei, 1997) cannot deal with gaps.

The alignment of tryptophanyl-tRNA synthetase sequences and tyrosyl-tRNA synthetase sequences is the same as the one reported in Diaz-Lazcoz et al. (1998). After the elimination of sites containing at least one gap, the sequences were 127 amino acids long.

These alignments, like all the other files used in the analysis, are available upon request.

The maximum parsimony criterion used in reconstructing the ancestral sequences was employed using the PAUP 3.1.1
program (Swofford, 1993). In order to reconstruct the ancestral sequences, the states for interior nodes option was used after building the specific topologies of the phylogenetic trees, selecting accelerated transformation (ACCTRAN) as the method for optimising characters (Swofford, 1993; Di Giulio, 2000b).

The reconstruction of the ancestral sequences by means of the maximum likelihood method was achieved using the ANCESTOR program of Zhang and Nei (1997). All the authors’ recommendations were followed in using this program.

The thermophily index (TI) that can be associated to any one protein sequence has already been defined (Di Giulio, 2000a). Briefly, it is defined by the expression:

\[
TI = \sum_{j=1}^{N} \frac{R_j}{N}
\]

where \(R_j\) is the value of the \(j\)th amino acid’s thermophily rank (Di Giulio, 2000a), and \(N\) is the total number of amino acids in the considered protein (Di Giulio, 2000a).

The optimal growth temperature \(T_{\text{opt}}\) values of the various organisms were taken from Jacobs and Gerstein (1960) and from Staley et al. (1984). In certain cases, especially for eukaryotes, these values were found by consulting the specialised literature.

3. Results

Fig. 1 shows the correlation between the optimal growth temperature of the various organisms and the thermophily index (TI) (Di Giulio, 2000a) for 84 amino acid sequences: 47 from the signal recognition particle 54 kDa (SRP54) and 37 from the SR subunit \(\alpha\). The regression line \(T_{\text{opt}} = -382.035 + 41.381TI\) was highly significant \((F = 41.14, df = 83, P < 10^{-4})\).

It is clear that if we estimated the TI value for the ancestral sequence of the last universal common ancestor (LUCA) from and on this sequence set (Fig. 1), then the assignment of this value to mesophilic or (hyper)thermophilic sequences would be difficult because the mesophilic sequences extend over a large range of the thermophily index (Fig. 1). Therefore, I have removed 29 points from the correlation in Fig. 1 and obtained the result shown in Fig. 2, in which there is a clear separation between the mesophilic and (hyper)thermophilic sequences. Therefore, the assignment of the TI value of the LUCA’s ancestral sequence (reconstructed from this sequence set (Figs. 2 and 3)) to the mesophiles or the (hyper)thermophiles should be less ambiguous in this case. This points removal can be justified by the fact that it does not in any way affect the estimate of the LUCA’s ancestral sequence, provided that the set of sequences thus obtained is representative of the spectrum of the three main lines of divergence, and in this case it is. Furthermore, using the variability present in the original data, for instance in Fig. 1, to calculate the interval of confidence (see legend to Tables 1 and 2) makes this removal virtually inoffensive.

In theory, there should be no need to use the sequences of two paralogous proteins to reconstruct the LUCA’s ancestral sequence but the sequences from a single protein might be sufficient. However, in practice this is not the case because the ANCESTOR program (Zhang and Nei, 1997) cannot work on rooted phylogenetic tree topologies and, thus, cannot estimate the sequence of the LUCA’s node. The use of paralogous proteins removes this limitation and therefore makes it possible to estimate the LUCA’s ancestral sequence, because the unrooted tree of the sequences of two paralogous proteins obviously contains two nodes for the LUCA corresponding to the deepest nodes of the set of sequences for every single orthologous protein.

The phylogenetic tree topologies that I have built are derived from that of the small subunit of ribosomal RNA (Maidak et al., 1997). For the set of sequences in Fig. 2, an example of these topologies is reported in Fig. 3 which
Fig. 3. The topology of one of the unrooted phylogenetic trees used in the analysis. The rooted topology for every single orthologous protein in the Bacteria domain is that of the ribosomal RNA of the small subunit (Maidak et al., 1997). To obtain the complete name of the species, see the web site mentioned in Section 2. The number ‘1’ at the end of the organisms’ names identifies the paralogous proteins in the α subunit, while the organisms with no number are those of the SRP54. See text for further information.
shows an unrooted tree in which the Eukarya domain is the Archaea’s sister domain. Moreover, the ancestral sequences were estimated using all three possible rootings of the tree of life that can be obtained from paralogous protein pairs.

Table 1 reports most of the information obtained from this analysis, i.e. the TI value for the reconstructed ancestral sequences of the ancestors we are interested in, and the estimate of the optimal growth temperature value that can be associated to these sequences (Table 1 and its legend).

I have carried out an equivalent analysis for the pair of paralogous proteins of the tryptophanyl- and tyrosyl-tRNA synthetases using a total of 49 sequences (Diaz-Lazcoz et al., 1998). Likewise for these sequences, a strong correlation is obtained \( F = 23.38, \text{df} = 48, \ P < 10^{-4} \) between the optimal growth temperature and the thermophily index (data not shown). The removal of 8 points from this correlation produces the regression shown in Fig. 4, whose points refer to 16 sequences of tryptophanyl-tRNA synthetase and SRP54.

![Fig. 4. Correlation between the optimal growth temperatures of the various organisms and the thermophily index for the pair of paralogous proteins of the tryptophanyl- and tyrosyl-tRNA synthetases. See text for further information.](image)
to 25 sequences of tyrosyl-tRNA synthetase. Finally, the
building of three phylogenetic trees with all the possible
rootings of the tree of life made it possible to reconstruct
the ancestral sequences (from the set of sequences in Fig. 4)
by means of both maximum parsimony and maximum like-
lihood and produced the temperature estimates reported in
Table 2.

### 4. Discussion

The analysis of the pairs of paralogous proteins of the
signal recognition particle clearly shows that the last univer-
sal common ancestor (LUCA) was a hyperthermophile
‘organism’ (Table 1). This is particularly true for the ances-
tral sequences estimated using maximum parsimony rather
than for those derived by means of maximum likelihood
(Table 1). Furthermore, this result is independent of how
the tree of life is rooted. All three possible rootings give rise
to a ‘hot’ LUCA (Table 1). Indeed, the set of sequences used
in the present paper (Figs. 2 and 3) make it difficult to
recover the alternative hypothesis of a mesophile LUCA.
This can only be obtained when, by preserving the identity
of the three domains, the sequences are ordered in such a
way that the thermophily index (TI) values of the sequences
in the phylogenetic tree topology (rooted in the Eukarya
domain) range from the lowest value, close to the LUCA
node, to the highest value, towards the less deep nodes (data
not shown).

The ancestors of the Bacteria and Archaea domains are
also hyperthermophiles (Table 1) while the ancestor of the
Eukarya domain seems to be a mesophile (Table 1), above
all if the TI values and hence the corresponding tempera-
tures are seen on Fig. 2 and not estimated by means of the
regression line, which raises these temperature values.

Overall, these observations are consistent with a large
quantity of data, suggestions and theories (Woese, 1987;
Achenbach-Richter et al., 1987; Wachtershauser, 1988,
However, the observation of a hyperthermophile LUCA clashes with the findings of an equivalent work (Galtier et al., 1999; see also Galtier, 2001) which found the LUCA to be a mesophile, although the data referred here agree with analyses of a similar nature (Di Giulio, 2000a,b). It is clear that only by perfecting the methods for reconstructing ancestral sequences, i.e. by better understanding molecular evolution, will we be able to give a ‘definitive’ answer to this question.

The observation of a non-mesophile LUCA is also confirmed for the pair of paralogous proteins of tryptophanyl- and tyrosyl-tRNA synthetases (Table 2), although in this case the LUCA seems to be more a thermophile than a hyperthermophile, at least for the estimates deriving from maximum likelihood (Table 2). (The anomalous behaviour of the Bacteria domain ancestor, which seems to be a mesophile (Table 2), can be explained by the fact that two sequences from hyperthermophiles (and in particular the sequence from *Aquifex aeolicus*) have a very low TI value (Fig. 4.).)

In order to transform the LUCA from a (hyper)thermophile to a mesophile, it is sufficient in this case to transfer the sequence from *Aquifex aeolicus* from the position near the LUCA’s node to positions close to the less deep nodes in the topology that sees the Bacteria domain as containing the root of the tree of life (data not shown). This occurs both for the estimates derived by maximum parsimony and those derived by maximum likelihood. In this case, therefore, the presence of hyperthermophile species close to the LUCA’s node is fundamental to determining whether or not it is a (hyper)thermophile.

The average accuracy (Zhang and Nei, 1997) of the LUCA’s ancestral sequences reconstructed using maximum likelihood (ML) has a mean value of 0.87 for the paralogous protein pair SRP54 and SR subunit α, and 0.76 for the tryptophanyl- and tyrosyl-tRNA synthetases. Therefore, the accuracy of the reconstruction for these deep nodes is relatively low. Furthermore, the comparison between the LUCA’s ancestral sequences reconstructed with maximum parsimony (MP) and those reconstructed using ML indicates that they are relatively dissimilar, with a mean difference in the identity percentage of 21% for those of the SRP54 and SR subunit α, and 30% for those of tryptophanyl- and tyrosyl-tRNA synthetases. This seems to indicate that the two methods grasp different aspects of molecular evolution. An example of this is reported in Fig. 5, which compares two of these sequences. These refer to the LUCA’s ancestral sequence, estimated using MP and ML, relative to the SR subunit α. The identity percentage between these sequences (Fig. 5) is 77.7% although the similarity percentage is 95.4% (Fig. 5). Overall, we can say that even if the sequences estimated by the two methods are relatively dissimilar, the conclusion that the LUCA was a thermophile or a hyperthermophile ‘organism’ is on the whole strengthened. This is because the difference between the ancestral sequences reconstructed by the two different methods seems to regard many aspects that should have been present in the true ancestral sequences.

Finally, whether or not the observation of a thermophile or hyperthermophile LUCA has any relevance in establishing that the origin of life took place at a high temperature is a question that has been comprehensively discussed elsewhere (Woese, 1987; Achenbach-Richter et al., 1987; Wächtershäuser, 1988, 1998; Pace, 1991; Wiegel and Adams, 1998; Arntienius et al., 1999; Galtier et al., 1999; Vogel, 1999; Di Giulio, 2000b; Nisbet and Sleep, 2001).

![Fig. 5. The ancestral sequences of the LUCA for the SR’s subunit α, estimated using maximum parsimony (MP) and maximum likelihood (ML). These sequences respectively refer to the LUCA’s nodes in the tree topology rooted in the Bacteria domain (Fig. 3). The star indicates that the same amino acid is present in the two sequences, the colon and the period respectively indicate similar or partially similar amino acids, while the white space indicates dissimilar amino acids.](image-url)
Acknowledgements

Part of this work was carried out at the Institute for Theoretical Physics of the University of California at Santa Barbara and was supported by the National Science Foundation under Grant no. PHY 99-07949.

References